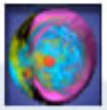
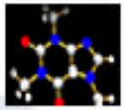
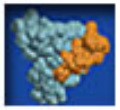



- PETASCALE DATA STORAGE INSTITUTE
  - 3 universities, 5 labs, G. Gibson, CMU, PI
- SciDAC @ Petascale storage issues
  - www.pdsi-scidac.org
  - Community building: ie. PDSW-SC07
  - APIs & standards: ie., Parallel NFS, POSIX
  - Failure data collection, analysis: ie., CFDR
  - Performance trace collection & benchmark publication
  - IT automation applied to HEC systems & problems
  - Novel mechanisms for core (esp. metadata, wide area)

Carnegie Mellon

center for information technology integration

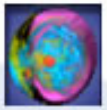
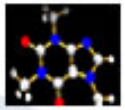
UNIVERSITY OF MICHIGAN

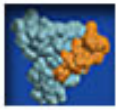
Sandia National Laboratories

Pacific Northwest National Laboratory
Operated by Battelle for the U.S. Department of Energy

Carnegie Mellon
Parallel Data Laboratory

pdsi

- Principle Petascale Storage issue is Scale
  - Up to Terabytes/sec bandwidth
  - Widely concurrent write sharing; non-aligned small strided
  - Trillions of files needing to do “ls -l”, “du -s”, backup
  - Billions of files in a directory
  - Millions of files creates and written per minute
  - Increasing need for brute force search
  - An order of magnitude or two more disks
  - Many more frequent failures, multiple failures
  - Operational staff costs not increasing
  - Weak programming for storage skills

Eg. POSIX Ext: Lazy I/O data integrity

- O_LAZY in *flags* argument to **open**(2)
- Requests lazy I/O data integrity
  - Allows filesystem to relax data coherency to improve performance for shared-write file
  - Writes may not be visible to other processes or clients until after **lazyio_propagate**(2), **fsync**(2), or **close**(2)
  - Reads may come from local cache (ignoring changes to file on backing storage) until **lazyio_synchronize**(2) is called
  - Does not provide synchronization across processes or nodes – program must use external synchronization (e.g., pthreads, XSI message queues, MPI) to coordinate

THE Open GROUP
Making standards work®

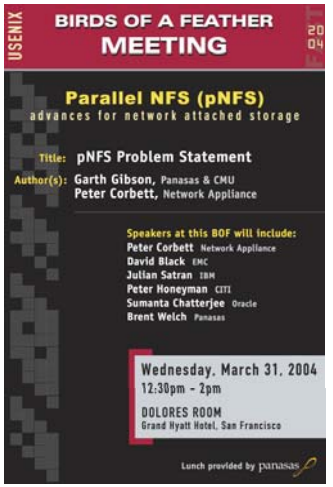
High End Computing Extensions Working Group

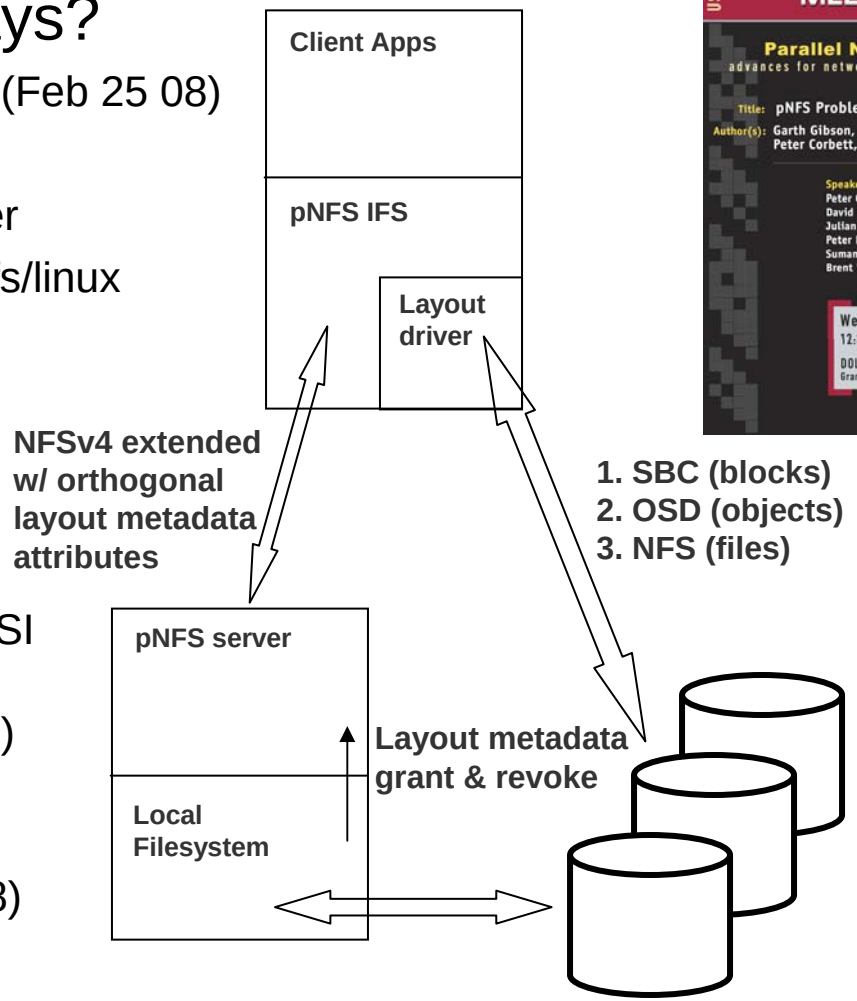
You are here: Platform Forum > HECEWG > Documents

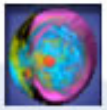
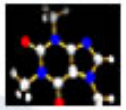
Created	Title (see details)	Version (+ implies others)	Formats (download)
17-Aug-2006	Evaluation Criteria for Proposed High End Computing Extensions to the POSIX I/O API	1.2 +	PDF
30-Jun-2006	Manpage - readdirplus	1	PDF
30-Jun-2006	Manpage - lockg (group lock)	1	PDF
30-Jun-2006	Manpage - sutoc (convert file handle to file descriptor)	1	PDF
30-Jun-2006	Manpage - NFSV4acls	1	PDF
30-Jun-2006	Manpage - openg (group open)		PDF
30-Jun-2006	Manpage - statlite and family of light weight stat calls	1	PDF
30-Jun-2006	Manpage - open (O_LAZY flags)	1	PDF
30-Jun-2006	POSIX I/O High Performance Extensions presentation Panasas SC05	1	PDF
30-Jun-2006	POSIX I/O High Performance Computing Extensions ASC SC05 presentation	1	PDF
30-Jun-2006	High End Computing Early Goals for extesions to POSIX I/O API	1	PDF
30-Jun-2006	A Business Case for Extensions to the POSIX I/O API for High End, Clustered, and Highly Concurrent Computing	1	PDF

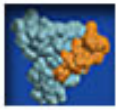
pNFS: Parallel File System Standards

- IETF NFSv4.1: soon 30 days?
  - draft-ietf-nfsv4-minorversion1-21.txt (Feb 25 08)
  - Includes pNFS, sessions
  - U.Mich/CITI impl'g Linux client/server
  - www.citi.umich.edu/projects/asci/pnfs/linux
- Three (or more) flavors of out-of-band metadata attributes:
  - FILES: NFS/ONCRPC/TCP/IP/GE for files built on subfiles NetApp, Sun, IBM, U.Mich/CITI
  - BLOCKS: SBC/FCP/FC or SBC/iSCSI for files built on blocks EMC (-pnfs-blocks-06.txt, Feb 25 08)
  - OBJECTS: OSD/iSCSI/TCP/IP/GE for files built on objects Panasas (-pnfs-obj-05.txt, Feb 25 08)





- PDSI Primary Early Emphasis:
  - Data Collection
    - Failure (next: Workload static/dyn)
    - Gather widely (LANL, NERSC, PNNL, ….)
    - Publish widely (CFDR w/ USENIX)

The computer failure data repository (CFDR)

With the growing scale of todays IT installations, component failure is becoming an ever larger problem. Yet, virtually no data on failures in real systems is publicly available, forcing researchers working on system reliability to base their work on anecdotes and back of the envelope calculations, rather than empirical data.

The computer failure data repository (CFDR) aims at accelerating research on system reliability by filling the nearly empty collection of public data with detailed failure data from a variety of large production systems.

Please join us, either by contributing data, downloading data, or joining our mailing lists.

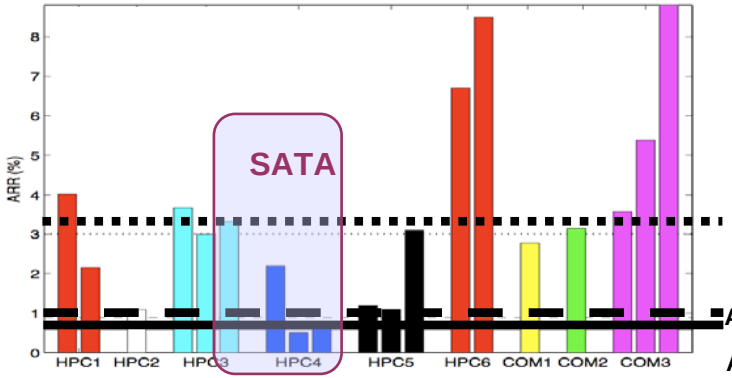
News

You are viewing a first draft of the CFDR. For feedback and comments please contact the moderators.

Available data

The table below provides an overview over the available data sets.

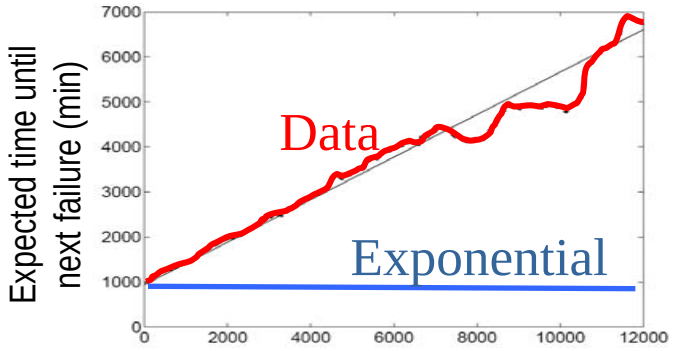
Name	Time period	System type	Type of data
LANL	Dec 96 - Nov 05	HPC clusters	The data covers node outages at 22 cluster systems at LANL, including a total of 4,750 nodes and 24,101 processors. Some job logs and error logs are available as well.
HPC1	Aug 01 - May 06	HPC cluster	The data covers hardware replacements at a 765 node cluster with more than 3,000 hard drives.
HPC2	Jan 04 - Jul 06	HPC cluster	Hard drive replacements in a 256 node cluster with 520 drives.
HPC3	Dec 05 - Nov 06	HPC cluster	Hard drive replacements observed in a 1,532-node HPC cluster with more than 14,000 drives.
PNNL	Nov 03 - Sep 07	HPC cluster	Hardware failures recorded on the MPP2 system (a 980 node HPC cluster) at PNNL.
COM1	May 2006	Internet services cluster	Hardware failures recorded by an internet service provider and drawing from multiple distributed sites.
COM2	Sep 04 - Apr 06	Internet services cluster	Warranty service log of hardware failures aggregating events in multiple distributed sites.
COM3	Jan 05 - Dec 05	Internet services cluster	Aggregate quarterly statistics of disk failures at a large external storage system.

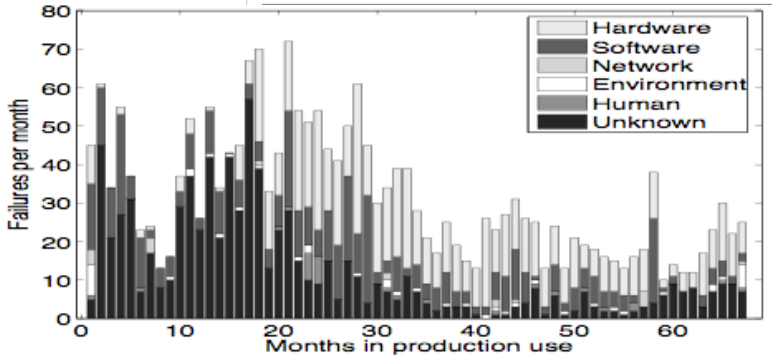


Data avrg = 3%

ARR = 0.88%

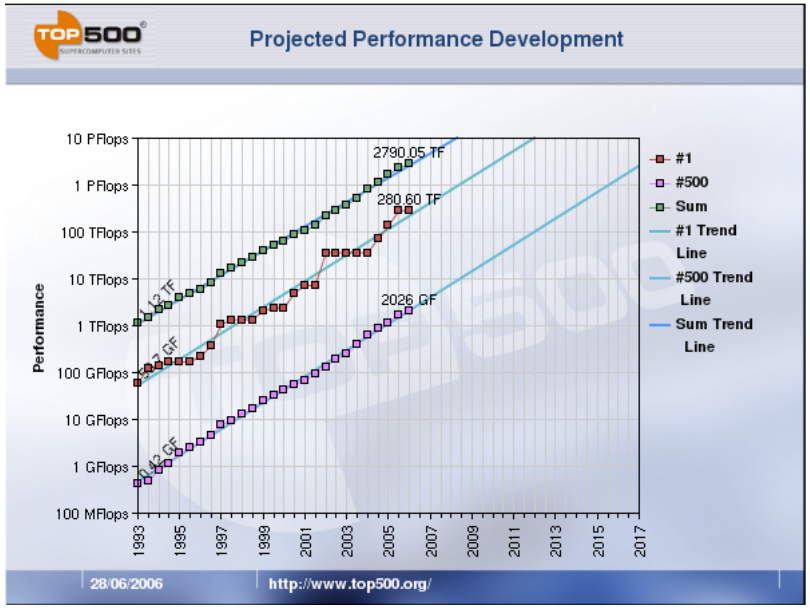
ARR = 0.58%

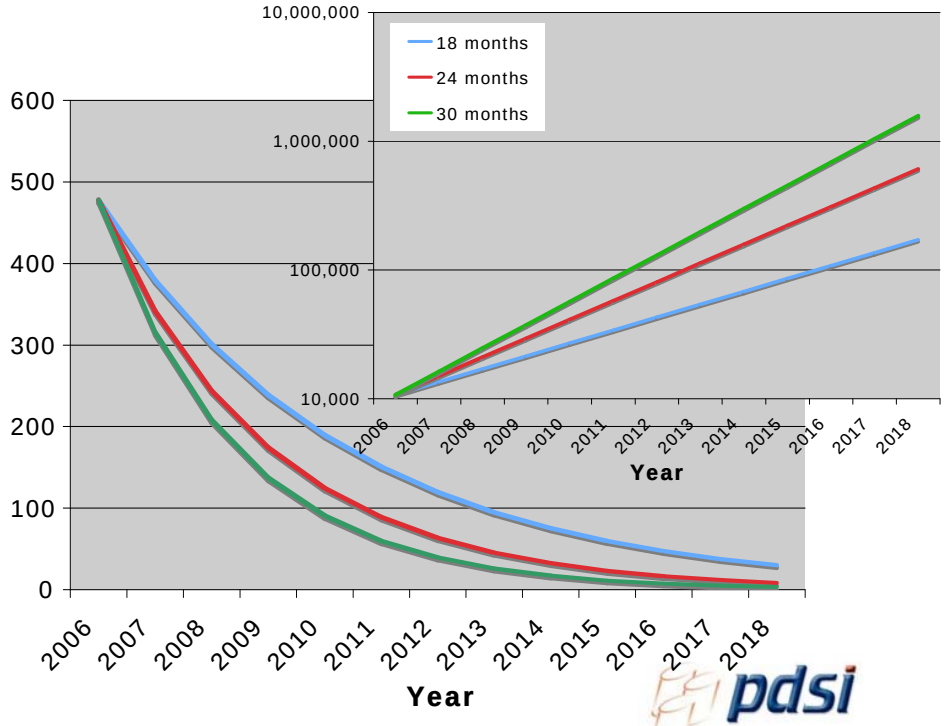




Peta/Exa-scale projections: more failures

- Continue top500.org annual 2X peak FLOPS
  - Talks: SciDAC07, ICPP07 Keynote, SEG (Oil&Gas), HECURA
- Cycle time flat; Cores/chip reaching for Moore's law
  - 2X cores per chip every 18-30 mos
- # sockets, 1/MTTI = failure rate up 25%-50% per year
  - Optimistic 0.1 failures/yr per chip (vs. LANL historic 0.25)



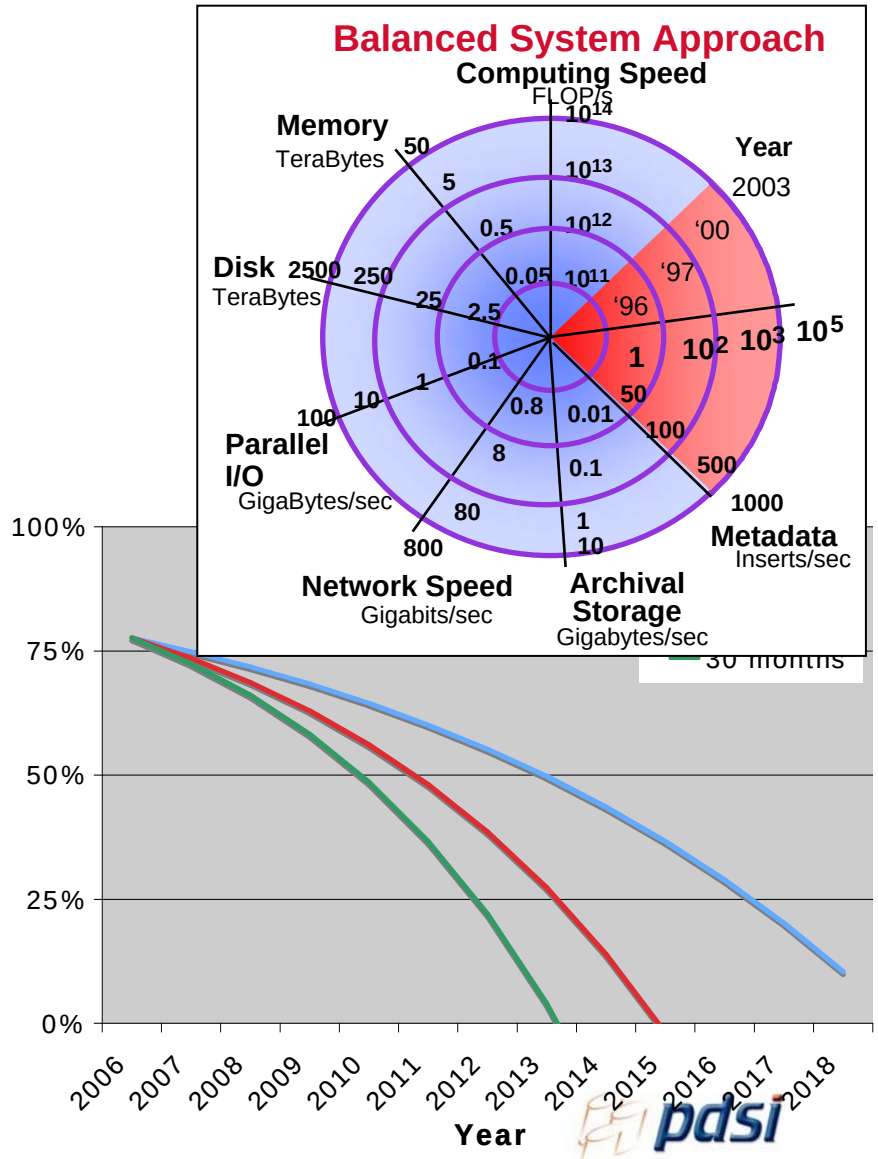


Carnegie Mellon
Parallel Data Laboratory

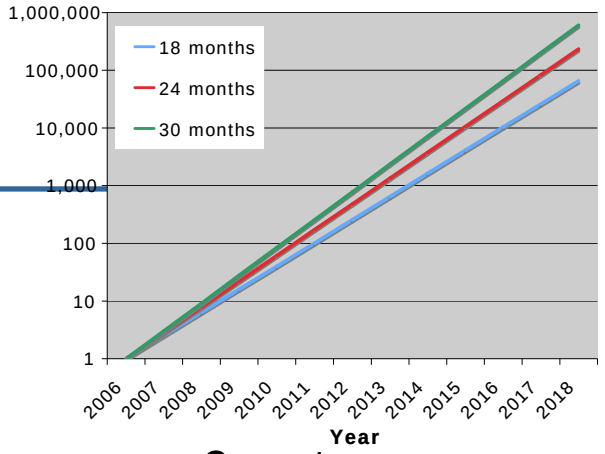
pdsi

Checkpointing failure tolerance faltering

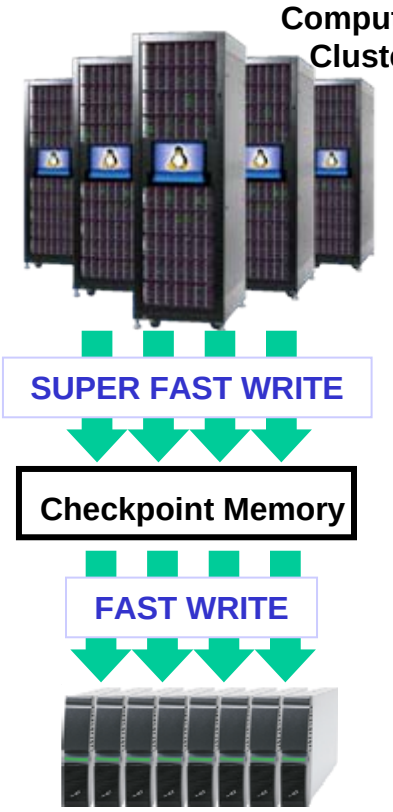
- Periodic (p) checkpoint (t)
- On failure, rollback & restart
- Balanced systems
  - Memory size tracks FLOPS
  - Disk speed tracks both
  - Checkpoint capture (t) constant
  - 1 - App util = t/p + p/(2*MTTI)
    $p^2 = 2*t*MTTI$
  - If MTTI was constant, app utilization would be too
  - But MTTI & app utilization drop
- Half machine gone soon and exascale era bleak

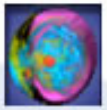
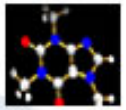


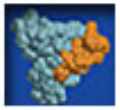
Fixes for Checkpoint/Restart



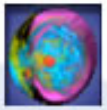
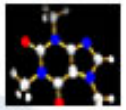
- Fix with more disk bandwidth?
  - Disk BW +20%/yr: Balance = +67% disks/yr
  - If MTTI drops, need +130% disks/yr !
- Smaller apps don't care
  - Constant HW & MTTI, so balance sufficient
- Compress memory image
  - 25%-50% smaller per byte per year fixes MTTI trend given balanced BW
- Process pairs: duplex all calculations
  - At 50% machine effectiveness, change to compute-thru-no-restart model
- Special purpose checkpoint devices
  - Fast memory to memory copy, offline to disk
  - Make copy "cheaper", say Flash

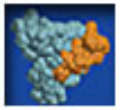


- PETASCALE DATA STORAGE WORKSHOP, SC07
  - Sponsored by PDSI with Program Committee:
  - Garth Gibson, Carnegie Mellon University & Panasas
  - Darrell Long, University of California, Santa Cruz
  - Peter Honeyman, University of Michigan, Ann Arbor, Center for Information Technology Integration
  - Gary Grider, Los Alamos National Lab
  - William Kramer, National Energy Research Scientific Computing Center, Lawrence Berkeley National Lab
  - Philip Roth, Oak Ridge National Lab
  - Evan Felix, Pacific Northwest National Lab
  - Lee Ward, Sandia National Lab

- PETASCALE DATA STORAGE WORKSHOP
  - Competitive extended abstract/paper selection
  - www.pdsi-scidac.org/SC07 for papers, presentations, posters as provided
- 22 submissions, 12 selected:
  - On Application-level Approaches to Avoiding TCP Throughput Collapse ….
  - pNFS/PVFS2 over Infiniband: Early Experiences
  - Integrated Systems Models for Reliability Petascale Storage Systems
  - Scalable Locking and Recovery for Network File Systems
  - Searching and Navigating Petabyte Scale File Systems Based on Facets
  - Scalable Directories for Shared File Systems
  - End-to-end performance management for scalable distributed storage
  - A Fast, Scalable, and Reliable Storage Service for Petabyte-scale ….
  - A Result-Data Offloading Service for HPC Centers
  - Characterizing the I/O Behavior of Scientific Applications on the Cray XT
  - A Universal Taxonomy for Categorizing Trace Frameworks
  - A Data Placement Service for Petascale Applications

PDSI Releases & Requirements

- This BOF is about community outreach
  - What resources we are already, and planning to make available to you
  - What problems and issues in Petascale Storage you might work on
- Part 1: this introduction
- Part 2: public releases
  - Bianca Schroeder, U. Toronto, Computer Failure Data Repository
  - Shobhit Dayal, CMU, fsstats: File Systems Statistics Database
  - Evan Felix, PNNL, PNNL released resources
  - Akbar Mokhtarani, NERSC, NERSC released resources
  - James Nunez, LANL, LANL released resources
  - Peter Honeyman, pNFS Linux open development
  - Ethan Miller, Ceph on sourceforge
- Part 3: petascale requirements
  - John Shalf, LBNL, “User Perspective on HPC I/O Requirements”
- Part 4: open discussion